

“Data is data” and the future of data browsing

Yaron Koren
March 22, 2013
SMWCon Spring 2013
New York, NY

Before we get to SMW, let's talk about data.

My new theory:

"Data is data".

The theory: the **structure** of a set of data says much more about how to interact with it than the data's **subject matter**.

In other words: if three different data sets have the same set of tables, column types, number of rows, etc., they can probably be viewed and edited with the same interface, even though the three are about:

- a children's TV show
- military information
- biotech research results

(For SMW, substitute "categories/classes" for tables, "property types" for column types, and "number of pages" (or "number of subobjects") for rows)

If this is true, what does it mean?

It means that there could potentially be generic software that takes in a set of data and constructs a reasonable interface for editing, browsing and visualizing that data, based only on the size and structure of that data.

(wow)

This could also be true regardless of the current storage of that data - database, spreadsheets, API, wiki, etc.

First, **browsing**.

If there is such a generic interface, what it would look like?

In other words, what is the ideal way to browse a set of data?

First: why do we browse data?

- To see the overall nature of the data
- To find patterns
- To find the set of entries that match some criteria
- To find a particular entry

Five basic approaches I know of:

- sortable table
- non-hierarchical drill-down
- hierarchical drill-down
- multi-pane
- faceted search

(These are my terms - are there more standard terms?)

Most of these have examples in the
MediaWiki/SMW world.

Sortable table of data:

Term extraction

This page lists term, n-gram and keyphrase extraction tools.

☐ Program Name	☐ Version	☐ Release date
AConCorde	0.4.2	23 December 2008
ExtPhr for Java	2006.07.28	
Kea	5.0	
Lucon	0.3.14	
Maui	1.2	
MyCAT	1.1.1b	31 July 2012
Ngram Statistics Package	1.21	
Poliqarp	1.3.6 (development version)	
TES	9.03	March 2009
TTC TermSuite	1.4	10 September 12
TextSTAT	2.9	19 March 2012
Topia	1.1.0	30 June 2009

Non-hierarchical drill-down: Semantic Drilldown extension

Organization

Click on one or more items below to narrow your results.

▼ Open or Free Statement?:

[no \(842\)](#) · [unknown \(32\)](#) · [yes \(245\)](#) · [yes - government \(63\)](#)

▼ License provider:

[ARR Copyright \(32\)](#) · [CC \(118\)](#) · [Creative Archive \(3\)](#) · [FSF \(5\)](#) · [None \(66\)](#) · [Open Content \(1\)](#)
· [custom \(56\)](#) · [public domain \(90\)](#) · [various \(30\)](#)

▼ License short name:

[CC BY \(47\)](#) · [CC BY-NC \(14\)](#) · [CC BY-NC-ND \(28\)](#) · [CC BY-NC-SA \(56\)](#) · [CC BY-ND \(4\)](#) · [CC BY-SA \(48\)](#) · [CC0 \(2\)](#) · [GNU FDL \(5\)](#) · [GNU GPL \(3\)](#) · [PD \(133\)](#) · [copyright \(814\)](#) · [custom \(46\)](#) · [various \(42\)](#)

▼ Organization Type:

[CMS privado \(1\)](#) · [Colectivo en internet \(1\)](#) · [Computer Science Journal \(1\)](#) · [Comunidad Organizada \(1\)](#) · [Consulting \(1\)](#) · [Education Center \(2\)](#) · [Enciclopedia \(1\)](#) · [Grupo de docentes \(2\)](#) · [Higher Education \(1\)](#) · [Higher Education Academy Subject Centre \(4\)](#) · [Higher Education Institution \(2\)](#) · [Independent \(1\)](#) · [Independent web-book \(1\)](#) · [LLC \(1\)](#) · [NGO \(2\)](#) · [None \(1\)](#) · [Online video show \(1\)](#) · [Privada \(1\)](#) · [Program to increase open textbook adoptions in](#)

The now-defunct “exhibit” query format, and its replacement, the “filtered” format, offer a more Javascript-y, slicker version of the same thing.

The screenshot shows a web interface for a children's calendar. At the top, there are three tabs: 'Calendar View' (highlighted in red), 'List View', and 'Search'. Below the tabs is the title 'Children and Young People 0-25 Calendar'.

There are two filter sections:

- Location:** A grid of checkboxes for various locations. 'London' and 'Salisbury' are checked.
- Children Activities:** A grid of checkboxes for different activities. 'Performing and arts' and 'Social and play' are checked.

Below the filters is a calendar for 'March 2013'. The calendar has a header with the month and year, and a 'Today' button. The days of the week are listed in the header. The calendar grid shows events for specific dates:

Sun	Mon	Tue	Wed	Thu	Fri	Sat
24	25	26	27	28	1	2
	Bridging Project Ages: 13-15 Trowbridge			Bridging Project Ages: 13-15 Salisbury		Kaya drum club Ages: 1-25 Trowbridge Zone Club Ages: 16+ Salisbury
3	4	5	6	7	8	9
	Bridging Project Ages: 13-15			Bridging Project Ages: 13-15		

Hierarchical drill-down:

- (1) category pages
- (2) #ask queries that display lists of pages

Category:Composers by nationality

From Wikipedia, the free encyclopedia

[Composers](#) by [nationality](#)



Wikimedia Commons has media related to: [Composers by country](#)

Subcategories

This category has the following 140 subcategories, out of 140 total.

► [Cabaret composers by nationality](#) (1 C)

*

► [Composers of classical music by nationality](#) (8 C)

► [Lists of composers by nationality](#) (61 P)

A

► [Afghan composers](#) (7 P)

F cont.

► [Flemish composers](#) (2 P)

► [Franco-Flemish composers](#) (73 P)

► [French composers](#) (8 C, 1,019 P)

G

► [Gabonese composers](#) (1 P)

► [Composers from Georgia \(country\)](#) (22

P)

N cont.

► [Nigerian composers](#) (7 P)

► [Norwegian composers](#) (2 C, 195 P)

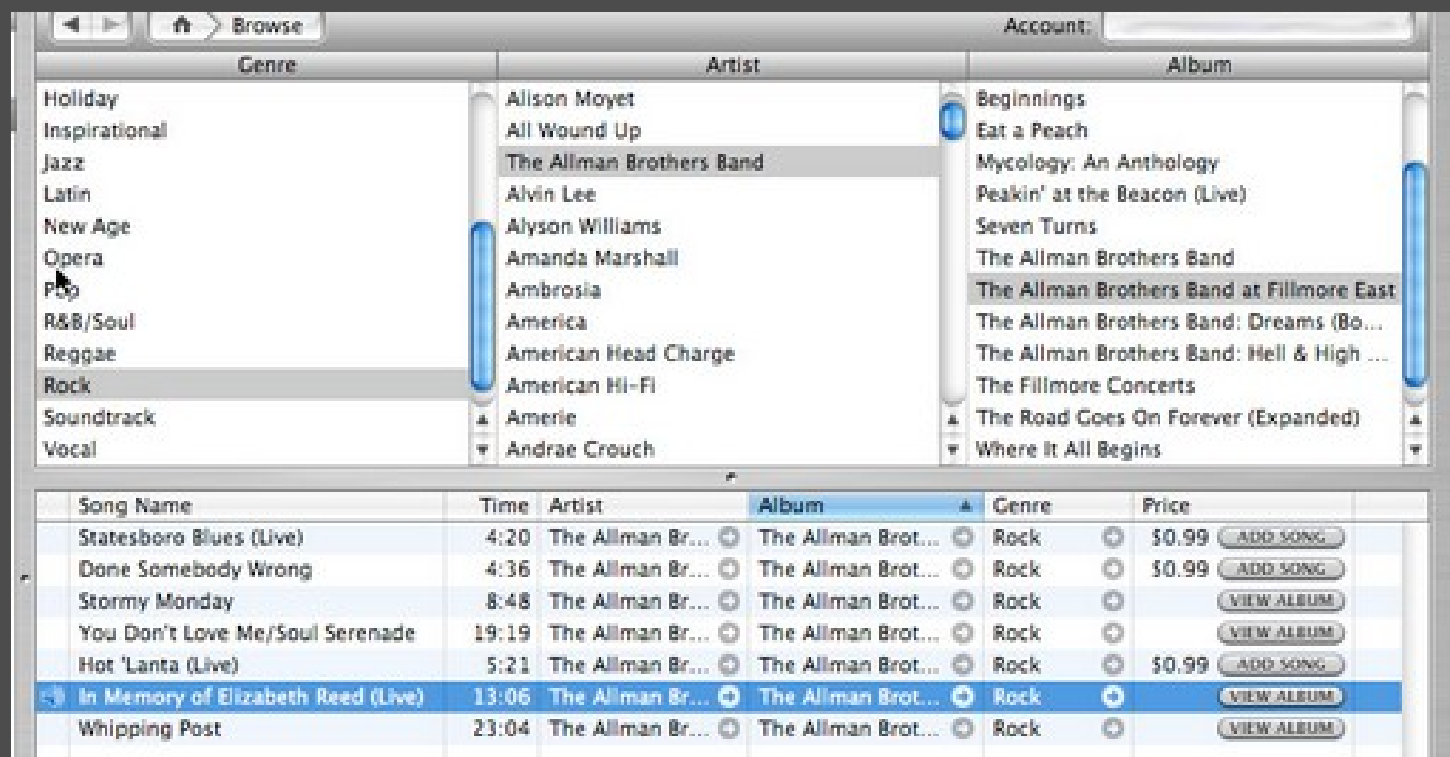
O

► [Ottoman composers](#) (5 P)

P

► [Pakistani composers](#) (1 C, 24 P)


Multi-pane: (a non-SMW example)



iTunes 4, from 2003.
(Now it's just a sortable table.)

Faceted search:

the "RunQuery" feature in Semantic Forms, the SolrStore and SOLRSearch extensions (previously also the SemanticQueryFormTool and Enhanced Retrieval extensions)



A screenshot of a web form for faceted search. It contains three input fields: a text box for 'Type of Menu You're looking for:', a text box for 'City:', and a dropdown menu for 'State:'. Below these fields is a button labeled 'Run query'.

Type of Menu You're looking for:

City:

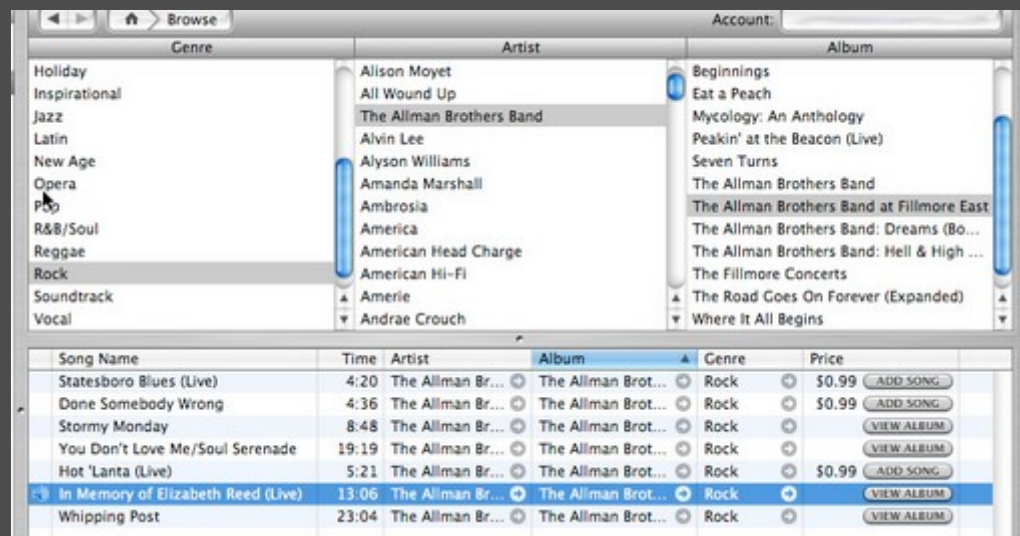
State: ▼

So which is the best?

Maybe there's no one answer.

It probably depends in part on the structure of the data – and on its size.

Example: A hierarchical drill-down, or multi-pane, interface, is good for hierarchical (1-to-n) data, right?



Genre > Artist > Album > Song

In SMW terms: hierarchical data is essentially any time a field/property can only hold a single value, not multiple values (no #arraymap).

Real-life data is rarely in a strict hierarchy - we use #arraymap a lot.

Genre > Artist > Album > Song

In the case of music, different songs on the same album can be by different artists, and of different genres.

(My iPod Nano (multi-pane interface) deals with this issue annoyingly: I can't play some albums all the way through, because it only combines songs together into "Album + Artist". Even Apple hasn't solved the data-browsing problem fully!)

Perhaps the right solution is to offer multiple interfaces, for users' different data needs?

Second, **visualization**.

If a table holds **geographical coordinates**, it should be displayable with a **map**.

If it holds a **date**, it should be displayable with an **outline, calendar or timeline** (depending on the date range).

If it holds a **number**, it should be **graphed**.

Can this be automated?

What does the "data is data" theory tell us about the future of data browsing in SMW?

1) Perhaps our interfaces should be smarter.

▼ Bioinformatics methods:

Ab-initio gene prediction (1) · Adapter Removal (software) (3) · Aligning (1) · **Alignment (53)** · Alignment Analysis (5) · Alignment viewer (6) · Alternative Splicing (1) · Annotation (11) · **Assembly (70)** · Assembly QC (3) · Assembly editing (1) · Assembly validation (3) · Assembly visualization (9) · Basecaller (12) · Basespace (1) · Biological Contextualization (2) · Biological Interpretation and Analysis of DNA Sequence Data (1) · Bisulfite SNP calling (1) · Bisulfite mapping (13) · Bloom filters (1) · Burrows-Wheeler (3) · ChIP-Seq (1) · ChIP-Seq analysis (2) · ChIP seq (2) · Chromatin motif finding (1) · Chromatogram management (1) · Chromatogram viewer (2) · Classification (1) · Clustering (4) · Clustering and alignment (1) · Collapsing Methods (1) · Colorspace (11) · Command line tool wrappers (2) · Community Analysis (1) · Comparative genomics (1) · Contaminant filtering (2) · Conversion (5) · Cost estimation (1) · Cufflinks (1) · Data compression (9) · Database (4) · Database interface (1) · Database submission preparation (2) · De-novo assembly (2) · De Bruijn graph (8) · Deduplication (1) · Differential expression (3) · Differentially expressed gene identification (8) · Differentially methylated regions identification and annotation (1) · Empirical Bayes (1) · Error correction (13) · Exome analysis (2) · Expectation Maximization (3) · Expression profiling (3) · FM-Index (4) · Filtering (11) · Format conversion (5) · GPU (5) · Gap extension (1) · Gene Set Testing (1) · Gene expression analysis (1) · Gene fusions discovery (1) · Gene ontology (2) · Gene ontology analysis (2) · Gene set enrichment (1) · General bioinformatics (1) · Genetic variation annotation (6) · Genome Alignment (3) · Genome Indexing (1) · Genome browser (7) · Genome wide association studies (1) · Genomic correlations (1) · Genomic overlaps (1) · Genomic region matching (2) · Genomics (1) · Genotyping (1) · Gibbs motif sample (1) · Graph reduction (1) · HLA typing (1) · Hadoop (8) · Haplotype reconstruction (4) · Hash Table Based (1) · Heatmaps (1) · Hidden Markov Model (12) · Hybrid assembly (2) · IGV (1) · Integrated Solution (5) · K-mer analysis (8) · LIMS (3) · Learning algorithm (1) · Localized reassembly/realignment (4) · MACS (1) · MCMC (1) · Machine Learning (2) · MapReduce (6) · **Mapping (113)** ·

Semantic Drilldown output, from SEQwiki

If we believe that "data is data", we can assume that this kind of display is never the ideal solution, and the software can do something more intelligent in this case.

2) A bigger deal: if one or more generic browsing applications are created, maybe we can switch to using those.

Perhaps SMW will one day get out of the data browsing business.

3) This may present a solution for mobile viewing of SMW data.

If the entire browsing apparatus consists of some data and some lightweight code around it, the whole thing could be run within **a mobile web browser**. (Most web browsers now contain a database.)

Finally, a side note about **editing**..

Editing a plain set of data, with no version history, is generally a bad idea.

However – our system could instead log a set of desired changes, and send it back to the central repository...

User Bob wants to
add [this row], and also
wants to change the
value of column 5 in
row 12 to “B”.



What could handle this message?

- A wiki
- Another content-management system
- A person

Questions/comments/complaints